A non-parametric hidden Markov model for climate state identification

M. F. Lambert, J. P. Whiting, A. V. Metcalfe

Hydrology and Earth System Sciences, 7(5), *652–667* (2003) © EGU

Hydrology & Earth System Sciences

A non-parametric hidden Markov model for climate state identification

Martin F. Lambert[1], Julian P. Whiting[1] and Andrew V. Metcalfe[2]

[1]Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, University of Adelaide, Adelaide 5005, Australia
[2]School of Applied Mathematics, University of Adelaide, Adelaide 5005, Australia

Email for corresponding author: mlambert@civeng.adelaide.edu.au

Abstract

Hidden Markov models (HMMs) can allow for varying wet and dry cycles in the climate without the need to simulate supplementary climate variables. The fitting of a parametric HMM relies upon assumptions for the state conditional distributions. It is shown that inappropriate assumptions about state conditional distributions can lead to biased estimates of state transition probabilities. An alternative non-parametric model with a hidden state structure that overcomes this problem is described. It is shown that a two-state non-parametric model produces accurate estimates of both transition probabilities and the state conditional distributions. The non-parametric model can be used directly or as a technique for identifying appropriate state conditional distributions to apply when fitting a parametric HMM. The non-parametric model is fitted to data from ten rainfall stations and four streamflow gauging stations at varying distances inland from the Pacific coast of Australia. Evidence for hydrological persistence, though not mathematical persistence, was identified in both rainfall and streamflow records, with the latter showing hidden states with longer sojourn times. Persistence appears to increase with distance from the coast.

Keywords: Hidden Markov models, non-parametric, two-state model, climate states, persistence, probability distributions

Introduction

Hidden Markov models (HMMs) have become popular tools for modelling dependent random variables in such diverse areas as speech processing (Juang and Rabiner, 1991), DNA recognition (Churchill, 1989) and rainfall occurrence (Zucchini and Guttorp, 1991). HMMs are based on a doubly stochastic process (Rabiner, 1989), in which an underlying stochastic process that develops as a Markov chain produces an unobservable ('hidden') state that can be inferred only through another set of stochastic processes. HMMs can be used to characterise observable signals as parametric random processes, the parameters of which can be determined in a precise manner (Rabiner, 1989).

Numerous authors have applied HMMs to stochastic hydrology, particularly in applications associated with climate variability. Zucchini and Guttorp (1991) applied a hidden Markov model to describe patterns of precipitation in space and time. In the construction of this model, the authors introduced unobserved climate states, which had different rainfall distributions associated with them. The transitions between these climate states were assumed to follow a Markov chain, with stationary transition probabilities.

Hughes and Guttorp (1994) described a nonhomogeneous hidden Markov model (NHMM) to relate broad scale atmospheric circulation patterns to local rainfall. The authors hypothesised unobserved, discrete-valued weather states, which classified atmospheric patterns into classes that are associated with particular precipitation patterns. In this model, transition probabilities between the hidden weather states depend on observable atmospheric data.

Thyer and Kuczera (2000) and Thyer (2000) use the concept of hidden climate states to provide a technique for the long-term simulation of hydroclimatic inputs for water resource planning, without the need to simulate supplementary climatic variables. The independent climate states, Wet (W) and Dry (D), which are assumed to have distinct rainfall distributions, are used in an attempt to explain varying wet and dry cycles in the Australian climate. At any given time, the distribution of observed random

variables in a HMM depends on the unobserved Markov chain, which governs the state of the process only through its value at that time (Ephraim and Merhav, 2002). These conditional distributions usually belong to a single parametric family (Robert *et al.*, 2000). When fitting these models to a random sample of data, it is unlikely that the parametric conditional state distribution of the population is known, and requires estimation. In their application of a two-state HMM to simulate annual rainfall recorded at three Australian coastal cities, Thyer and Kuczera (2000) assumed that rainfall in both climates was consistent with random draws from Gaussian distributions. At sites in which annual rainfall is skewed, it is unlikely that this model assumption will still be suitable.

In this paper, an innovative non-parametric HMM (NPHMM) is described which avoids assumptions about the distribution of the observed process in each state. Therefore, the estimation of transition parameters is not compromised by a conflicting requirement that the observed marginal distribution be a mixture of unrealistic conditional state distributions.

Parametric two-state HMM

BACKGROUND TO MODEL

HMMs can be described by a pair of discrete-time stochastic processes $\{(x_t, y_t)\}$. In this model, suppose that x_t is a finite-state Markov chain, described at any time as being in one of a set of k distinct states, $\{s^1, s^2, ..., s^k\}$. The set of one step transition probabilities, $P = \{p_{ij}\}$, which characterise changes in state is defined as

$$p_{ij} = P\left(x_t = s^j \middle| x_{t-1} = s^i\right) \; 1 \le i, j \le k \tag{1}$$

where

$$\sum_j p_{ij} = 1 \quad \text{for all } i \tag{2}$$

The unobserved process x_t is referred to as being in state j $(1 \le j \le k)$ at time t by the notation s_t^j. Given s_t^j, y_t is the observation at time t, with a probability distribution that depends only on the underlying state, formally defined by the following assumption:

$$P\left(y_t \middle| X_t, Y_{t-1}\right) = P\left(y_t \middle| x_t\right) \tag{3}$$

where Y_{t-1} is the sequence of observations from time 1 to time t–1, $\{y_1, y_2, ..., y_{t-1}\}$, and similarly for X_t. The observed process may be either discrete valued or continuous and is described as conditionally independent random variables.

The distribution of y_t without any conditioning state is the marginal distribution. It is estimated by the distribution of all the observations, without regard to their order of occurrence.

This paper will focus on two-state HMMs. The states are referred to as Wet (W) and Dry (D), reflecting the hypothesis of distinct, but not directly observable, climate states that influence rainfall.

HMM DEGENERATING TO A MIXTURE DISTRIBTUION

A mixture of two normal distributions can often closely approximate a skewed marginal distribution. A HMM includes a mixture as the special case of state transitions that are independent of the current state. That is

$$P_{DW} = P_{WW} \text{ and } P_{WD} = P_{DD} \tag{4}$$

where P_{DW} is the transition probability from a Dry to a Wet state, and so on. This condition is equivalent to

$$P_{DW} + P_{WD} = 1 \tag{5}$$

The proof that Eqn. (4) implies Eqn. (5) follows:

Let P_{dry} and P_{wet} be the proportion of time that the HMM spends in Dry and Wet states respectively.

Then

$$P_{dry} + P_{wet} = 1 \tag{6}$$

The stationary probability vector for a two-state HMM satisfies

$$\begin{bmatrix} P_{wet} & P_{dry} \end{bmatrix} \times \begin{bmatrix} P_{WW} & P_{WD} \\ P_{DW} & P_{DD} \end{bmatrix} = \begin{bmatrix} P_{wet} & P_{dry} \end{bmatrix} \tag{7}$$

and so

$$P_{wet} = \frac{P_{DW}}{1 + P_{DW} - P_{WW}} \tag{8}$$

If $P_{DW} = P_{WW}$, then

$$P_{wet} = P_{DW} = P_{WW} \tag{9}$$

Similarly

$$P_{dry} = P_{WD} = P_{DD} \tag{10}$$

and hence $P_{DW} + P_{WD} = 1$.

Conversely, if Eqn. (5) holds, the transition matrix must

have the form

$$\begin{bmatrix} p & (1-p) \\ p & (1-p) \end{bmatrix} \tag{11}$$

where $p = P_{WD} = P_{DD} = P_{dry}$

If state transitions are independent of the current state, the construct of climate states is redundant and the model can be thought of as merely a mixture. If there is a tendency for the model to persist in either state, the sum of P_{WD} and P_{DW} will be less than 1, whereas a sum greater than 1 would correspond to a tendency to fluctuate between states.

A Bayesian credibility interval for the sum of P_{WD} and P_{DW} can be used as an indicator of persistence. If the upper limit of the 90% credibility interval is less than 1, evidence of hydrological persistence can be claimed. However, if the interval includes 1 there is no convincing evidence to dismiss the possibility that the HMM is merely a mixture.

INAPPROPRIATE GAUSSIAN ASSUMPTION BIASES TRANSITION PROBABILITIES

A mixture of two Gaussian distributions can approximate a skewed marginal distribution, even if it has arisen as a mixture of two skewed conditional state distributions. However, the mixing proportions are likely to be different. A biased estimate of mixing proportions will lead to biased estimates of transition probabilities through Eqn. (8). This could result in missing statistically significant evidence for the existence of climate states, or possible incorrect claims for such evidence. The advantage of the non-parametric approach to be proposed later is that this possible source of bias is avoided. Another approach would be to adopt a more general family of distributions. However, the NP approach is still useful as it can be used to identify suitable probability distributions.

PARAMETER ESTIMATION

In the fitting of the HMM, it is necessary to estimate the unknown model parameters, denoted by the vector θ, from the observed time series. In the parametric case, θ consists of the parameters of the assumed probability distribution and transition probabilities. In the non-parametric case, θ includes a single transition probability and related parameters. Parameter estimates are described by the posterior distribution of θ, conditional on the entire set of observations Y_n, denoted by $p(\theta|Y_n)$, which can be represented through Bayes' Theorem as:

$$p(\theta|Y_n) = \frac{p(Y_n|\theta)p(\theta)}{p(Y_n)} \tag{12}$$

In this relationship, $p(Y_n)$ is the marginal probability $p(Y_n) = \int p(Y_n|\theta)p(\theta)d\theta$, $p(\theta)$ is the prior distribution of θ and $p(Y_n|\theta)$ is known as the likelihood distribution. The likelihood is central to the implementation of both the HMM and the NPHMM and can be expressed as:

$$p(Y_n|\theta) = p(y_1|\theta)\prod_{t=2}^{n} p(y_t|Y_{t-1},\theta) \tag{13}$$

where Y_{t-1} represents all the observations up to time $(t-1)$. Chib (1995) outlines an iterative procedure to evaluate the term $p(y_t|Y_{t-1},\theta)$ for a two-state HMM, which exploits the Markovian state dependence and proceeds in the following three steps. In these steps, s_t^i is representative of the condition $x_t = s^i$.

$$p(s_t^i|Y_{t-1},\theta) = \sum_j p(s_t^i|s_{t-1}^j,\theta)p(s_{t-1}^j|Y_{t-1},\theta) \tag{14a}$$

$$p(y_t|Y_{t-1},\theta) = \sum_j p(y_t|s_t^j,\theta)p(s_t^j|Y_{t-1},\theta) \tag{14b}$$

$$p(s_t^i|Y_t,\theta) \propto p(y_t|s_t^i,\theta)p(s_t^i|Y_{t-1},\theta) \tag{14c}$$

In this procedure, the probability density $p(y_t|s_t^j,\theta)$ is the likelihood of observing y_t given the climate at time t is in the j^{th} state, and $p(s_t^i|s_{t-1}^j,\theta)$ denotes stationary transition probabilities. The iteration of these three steps over all observations evaluates the likelihood of the observed time series.

In the HMM, it is not possible to derive analytically an expression for the posterior distribution $p(\theta|Y_n)$. When it is not possible to evaluate explicitly the posterior distribution, numerical integration or analytical approximation techniques are often required (Brooks, 1998). The Monte Carlo Markov Chain (MCMC) method provides an alternative, through the construction of aperiodic and irreducible Markov chains, which have stationary distributions that approximate the posterior distribution of interest. By running such chains for long enough, simulated variables can be treated as samples from the posterior distribution, and can be used for summarising important features of the posterior (Brooks, 1998). The Metropolis algorithm (Metropolis *et al.*, 1953; see also Kuczera and Parent, 1998) is perhaps the broadest implementation of these methods and is used in this paper to derive the posterior distributions of unknown model parameters.

Non-parametric two-state HMM

BACKGROUND TO MODEL

A non-parametric (NP) HMM has been developed as an alternative to the parametric two-state HMM described in

the previous section. This model has been termed 'non-parametric', as no assumptions concerning the underlying state distributions are made. The procedure used to fit this model to a known data series is summarised here:

Let $\{y_1, y_2, ..., y_n\}$ represent the observed data in time order, with y_t being the datum at time t, where $1 \le t \le n$. Now let $\{y\}$ be the data sorted into ascending order. A transform into the [0,1] interval is defined by $y_\tau \rightarrow \frac{m}{n+1}$ where m is the rank of y_τ when the observations are sorted. Define $u_\tau = \frac{m}{n+1}$ then $\{u_1, u_2, ..., u_n\}$ is the time series transformed into the [0,1] interval. The hidden model states will be identified within this transformed time series. As the sorting procedure employed to generate the transformed time series can be undertaken on both discrete and continuous variables, the non-parametric HMM can be fitted, without modification, to time series of either type.

By using the same model states that were used in the two-state parametric HMM, it is assumed that a value u_τ has arisen from either a Wet state (W) or a Dry state (D) and that the higher values of u_τ are more likely to be from the former. Therefore, for a given value of u_t, $1 \le t \le n$, the probability of that value having arisen from a Wet state is complementary to the probability of that value having arisen from a Dry state, i.e.

$$P(W|u_t) + P(D|u_t) = 1 \qquad (15)$$

Transitions between the two states are defined in the same way as for the parametric HMM. The states of the transformed series can be represented geometrically by the partition of a unit square that has u_t on the horizontal axis and $P(s_t^j | u_t, \theta)$ on the vertical axis (where $j = W, D$), giving a height of unity. The parameter vector θ has been included as a conditional vector of unknown quantities, which includes the HMM transition probabilities and a relationship governing the partition between the two states. This partition divides the square in such a way that the probabilities of u_t having been generated from either state are defined. Figure 1 illustrates the square structure of the NPHMM, with a partition curve separating the Wet and Dry states.

PARTITIONING THE SQUARE

The partition of the square on which the NPHMM is based can take a variety of shapes dependent upon the underlying state distributions. At one extreme, if the distributions are well separated, the partition will be a vertical line with areas either side corresponding to the proportion of the total from each distribution. To illustrate this, data from a simulation of 10 000 draws from Wet state $N(2000, 200^2)$ and Dry state $N(1000, 200^2)$ were put in ascending order. The complimentary histograms in Fig. 2a show the proportion of data from the Dry and Wet distributions for u_t bin widths of 0.01.

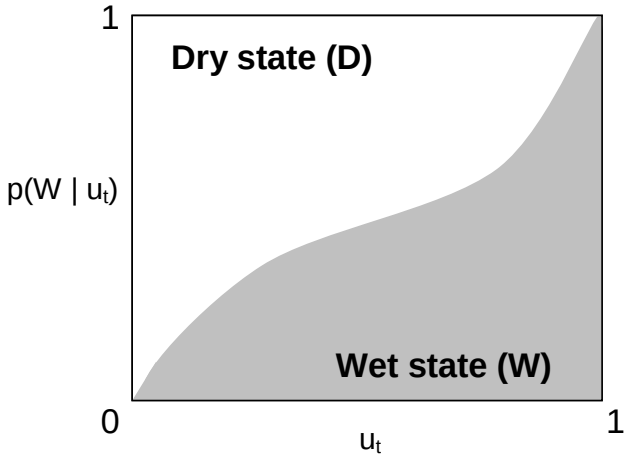

Fig. 1. *An example of the separation of the NPHMM unit square into Wet and Dry states*

At the other extreme of conditional distributions, the partition will be horizontal, with the areas either side of the partition corresponding to the proportions of the data from either distribution. As an illustration, data from a simulation of 10 000 draws from a Wet state and a Dry state that were both $N(2000, 200^2)$ were put in ascending order. Figure 2b shows complimentary histograms with the proportion of data from both distributions for u_t bin widths of 0.01.

To illustrate a partition that is between the horizontal and the vertical, data from a simulation of 10 000 draws from Wet state $N(2000, 200^2)$ and Dry state $N(1500, 200^2)$ were put in ascending order. Complimentary histograms in Fig. 2c show the proportion of data from Wet and Dry distributions for u_t bin widths of 0.01.

APPROXIMATING THE NON-PARAMETRIC DIVISION

The division between the Wet and Dry histograms in Figs. 2a-c can be characterised by a continuous curve. However, rather than specify any functional form, a continuous division is effected by a discrete number of contiguous line segments. The approximation of a curve of partition by three segments, constrained at points (0,0) and (1,1), will be determined by the coordinates of two points P_1 and P_2 as shown in Fig. 3. In the application of this model, a maximum likelihood procedure can determine the location of these two points, under the constraint that the coordinates of P_2 are greater than P_1. In this way, a Wet state is identified as lying below these segments.

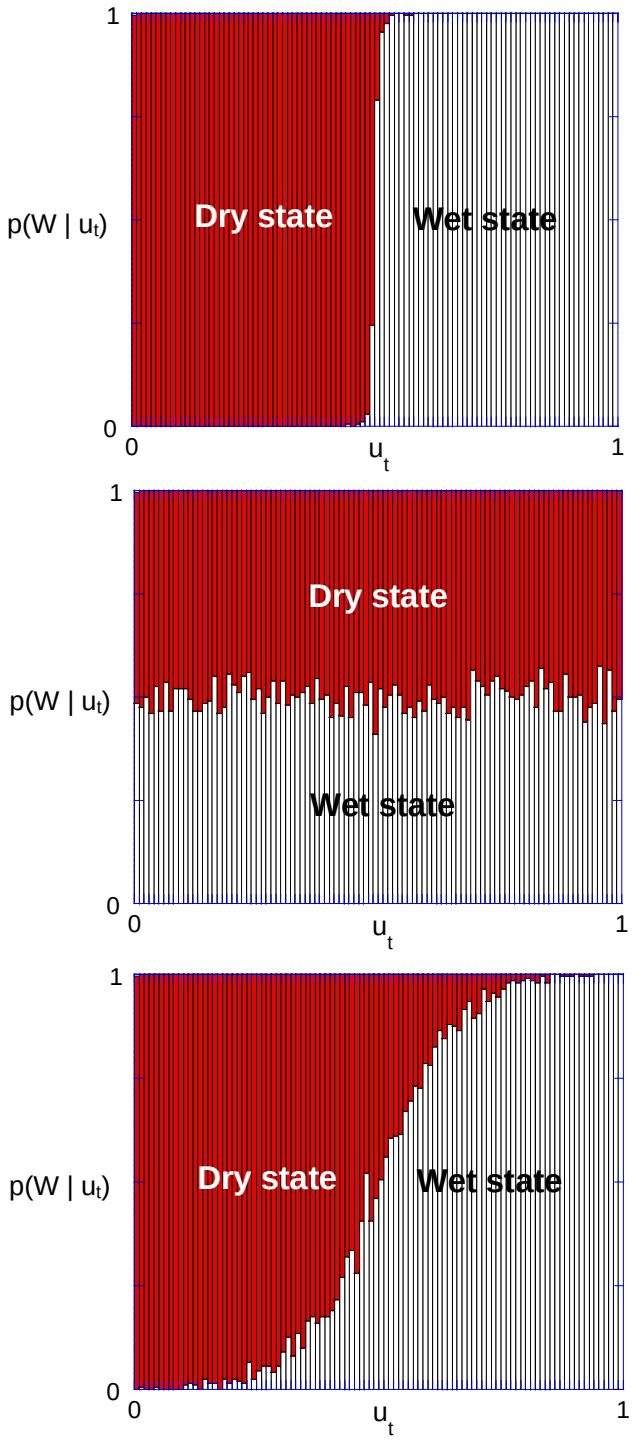

Fig. 2. *The estimated separation of NPHMM unit squares for samples from three simulated Markov chains with Wet and Dry distributions that are (a) well separated, (b) identical, (c) overlapping*

It is possible to generalise the NPHMM to include more than two states. In the model illustrated here, two independent states are defined by one partition. If this framework is extended to allow for three model states, two

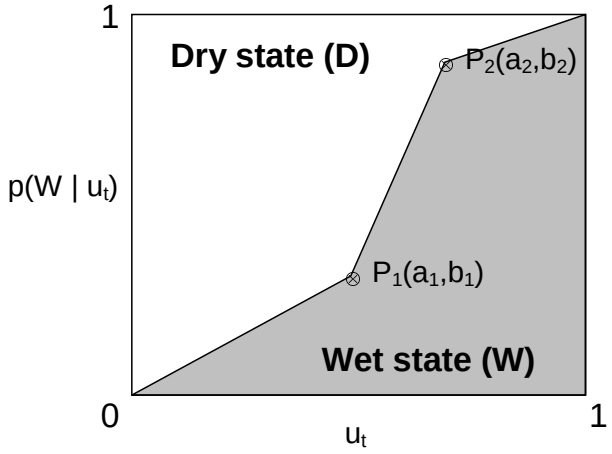

Fig. 3. *A two-point division of a two-state NPHMM unit square*

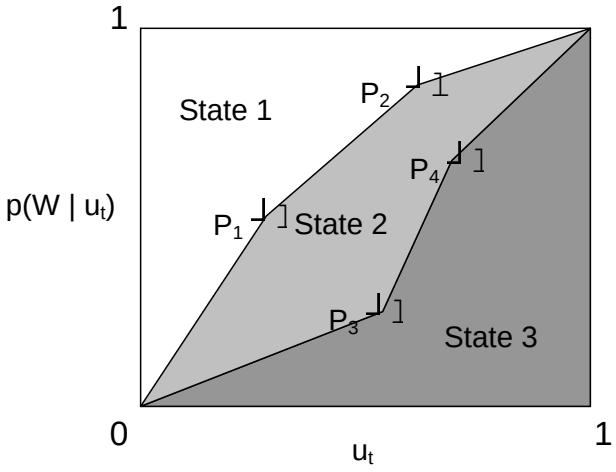

Fig. 4. *A four-point division of a three-state NPHMM unit square*

partitions need to be identified, as shown in Fig. 4. By approximating both partitions with three contiguous line segments, the locations of four points need to be identified. It is a straightforward procedure to extend the model further to allow for more than three states.

An alternative two-state model structure is to approximate the partition curve by a greater number of line segments. For example, ten contiguous lines will be separated by nine points, which can be equally spaced along the horizontal axis. A maximum likelihood procedure can determine the vertical coordinates of these nine points, under the constraint that the height of each point above the horizontal axis is not less than the height of the preceding point. This possible 'nine-point' model is shown in Fig. 5. As in Fig. 3, the ten line segments in this model are constrained at points (0,0) and (1,1).

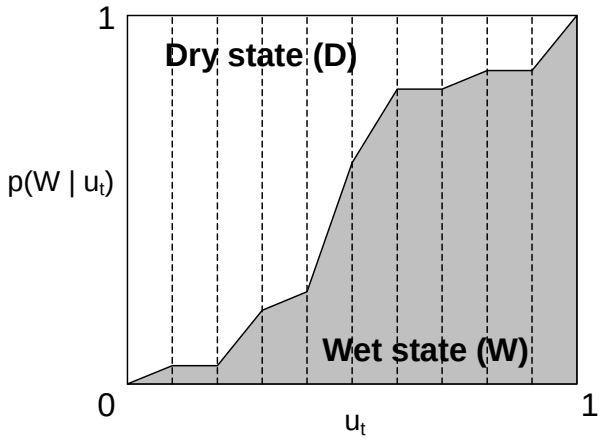

Fig. 5. *A nine-point division of a two-state NPHMM unit square*

In the 'two-point' model and the 'nine-point' model, the segment of a vertical line through the point u_t, between the horizontal axis and the partition curve is $P(W|u_t,\theta)$. The length of this line above the partition is therefore $1-P(W|u_t,\theta)$, or $P(D|u_t,\theta)$. The parameter vector θ includes the locations of the points that define the partition curve. In this paper however, the 'two-point' NPHMM will be used exclusively.

NON-PARAMETRIC LIKELIHOOD FUNCTION

In the Chib (1995) likelihood function (Eqn. 14a-c), the probability density function (pdf) for the observation in a given state is represented by $p(y_t|s_t^j,\theta)$, where $j=W,D$ in this case. This probability can be related to the pdf of the NPHMM via the uniform transformation. The non-parametric HMM pdf in state *j* can be represented as $p(u_t|s_t^j,\theta)$. Now, from Bayes' theorem, this probability is expressed as:

$$p(u_t|s_t^j,\theta)=\frac{P(s_t^j|u_t,\theta)\times p(u_t|\theta)}{P(s_t^j|\theta)} \qquad (16)$$

Here, $P(s_t^j|u_t,\theta)$ is the probability of u_t being in state *j* at time *t*, which is the height of the partition at that point, as defined earlier. Also, $p(u_t|\theta)$ is the marginal pdf of the value u_t, which will be equal to unity since the transformed series is mapped to a uniform distribution. Furthermore, $P(s_t^j|\theta)$ is the marginal probability of each state (the steady-state probabilities P_{wet} and P_{dry}) given the shape of the partition curve. These marginal probabilities are equal to the proportions of the area of the uniform distribution described as Wet or Dry by the partition and are related to the transition probabilities by the relationships

$$P_{wet}=\frac{P_{DW}}{P_{DW}+P_{WD}} \qquad (17a)$$

and

$$P_{dry}=\frac{P_{WD}}{P_{DW}+P_{WD}} \qquad (17b)$$

As a result, when fitting the two-state NPHMM to a data set, only one of the transition probabilities needs to be estimated, the other being related through the location of the partition curve. Consequently, the partition can be scaled through Bayes' theorem to form a pdf for use in the likelihood function iterations (Eqns. 14a-c). Thus the NP likelihood function can be calculated and, given a suitable prior distribution, MCMC procedures can be utilised as with the parametric HMM to estimate the posterior distribution of unknown parameters, $p(\theta|U_n)$. In this paper, uniform prior distributions over a [0,1] interval for both P_{WD} and P_{DW}, and for the coordinates of the partition within the unit square, have been assumed.

IDENTIFICATION AND ESTIMATION OF STATE DISTRIBUTIONS

Following the identification of the location of the partition, a Monte Carlo sampling procedure is used to produce estimates of the conditional state distributions:

i. A uniform random number is generated, relating to a value u_τ and its corresponding position on the horizontal axis the non-parametric HMM square.
ii. A corresponding value (y_τ) from the original time series is interpolated from the transformed value u_τ.
iii. A second uniform random number (p_τ) is generated and yields a distance along the orthogonal line through u_τ on the square.
iv. If p_τ lies above the partition, then the value of y_τ will be assigned to the Dry state distribution, and vice versa for the Wet state.

Estimates of the two underlying state distributions can then be obtained by repeating this sampling procedure multiple times (100 000 repetitions in the results shown in this paper). These distributions are guided only by the location of the partition line, which is identified through a maximum likelihood procedure. In this way, no assumptions about the underlying state distributions in the two climate states have been made.

Application of the NPHMM to simulated data

The NPHMM is fitted to time series that are simulated from various distributions for a length of 1000. This length is

chosen as many continuous monthly rainfall series have observations of between 80 and 100 years, with the former equating to approximately 1000 values. It is useful to be familiar with the ability of the model to identify known model parameters from simulated time series before interpreting results from fitting the model to observed data.

MARKOV CHAIN SIMULATED FROM TWO GAUSSIAN DISTRIBUTIONS

A two-state Markov chain of length 1000 was simulated with transition probabilities P_{WD} = 0.25 and P_{DW} = 0.15. Random samples in a Wet state were simulated from $N(1500,200^2)$, with a Dry state being sampled from $N(1000,200^2)$. Following the fitting of the two-point NPHMM to this Markov chain, estimates of the Wet and Dry distributions were obtained with means 1482.7 and 1002.8 and standard deviations 205.6 and 210.1 respectively. These estimated state distributions are compared to the original simulated distributions in Fig. 6, and show a satisfactory reproduction of the original data. From fitting the NPHMM, the posterior distributions of the two transition probabilities, P_{WD} and P_{DW}, have means 0.237 and 0.126. From these distributions, 90% credibility intervals are [0.189, 0.283] and [0.098, 0.157] respectively, both of which include the values used in the simulation. For comparison, the Gaussian HMM was fitted and the corresponding credibility intervals for P_{WD} and P_{DW} are [0.195, 0.291] and [0.105, 0.163]. These intervals are unbiased and slightly narrower, as expected when assumptions of parametric models are valid.

MARKOV CHAIN SIMULATED FROM TWO LOGNORMAL DISTRIBUTIONS

For this simulation, a two-state Markov chain was simulated with transition probabilities P_{WD} = 0.30 and P_{WD} = 0.10 for a length 1000. Random samples for a Wet state were first drawn from $N(7.5,0.25^2)$ with random draws in a Dry state from $N(6.8,0.22^2)$, with exponentials of these samples then taken.

Following the fitting of the two-point NPHMM to this Markov chain, estimates of the two underlying state distributions are obtained. Taking logarithms of these estimates produces Wet and Dry distributions that have means 7.498 and 6.812 and standard deviations 0.242 and 0.226 respectively. These parameter estimates are close to the known values, and the estimated distributions are compared to the original simulated distributions in Fig. 7. From fitting the NPHMM, the posterior distributions of the two transition probabilities, P_{WD} and P_{DW}, have means 0.306 and 0.107 respectively. From these distributions, 90% credibility intervals are [0.244, 0.371] and [0.088, 0.128], both of which include the values used in the simulation. For comparison, the Gaussian HMM was fitted and the corresponding credibility intervals for P_{WD} and P_{DW} are [0.291, 0.408] and [0.115, 0.163] respectively. These intervals are biased, as expected when the assumptions of

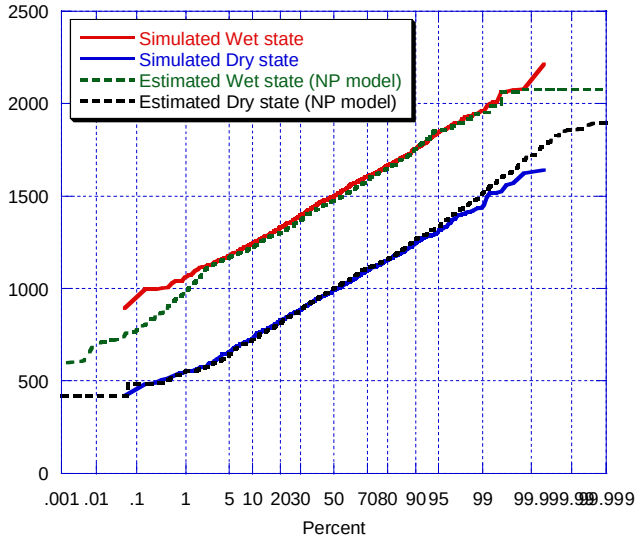

Fig. 6. *Simulated state conditional Gaussian distributions of a variable (solid line) compared with the estimated distributions (dotted lines) from the NPHMM*

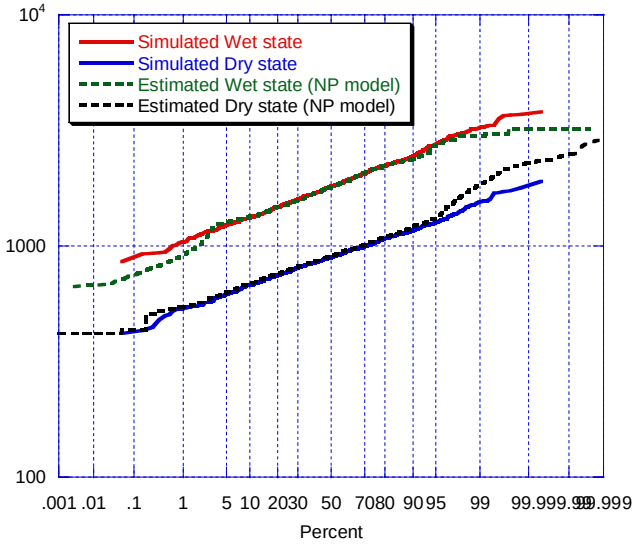

Fig. 7. *Simulated state conditional lognormal distributions of a variable (solid line) compared with the estimated distributions (dotted lines) from the NPHMM*

parametric models are infringed, and the latter excludes the value of 0.10 used in the original simulation.

MARKOV CHAIN SIMULATED FROM ONE GAUSSIAN AND ONE LOGNORMAL DISTRIBUTION

In this example, a two-state Markov chain of length 1000 was constructed with transition probabilities $P_{WD} = 0.25$ and $P_{DW} = 0.15$. Random samples in a Dry state were drawn from $N(1000,200^2)$ with Wet state samples being the exponentials of random draws from $N(7.4,0.3^2)$.

After fitting the two-point NPHMM to this Markov chain, an estimated Dry state was distributed with a mean 1020.7 and standard deviation 202.5, whereas the logarithms of the estimated Wet state were distributed with a mean 7.394 and standard deviation 0.290, corresponding to a sample mean of 1696.7 and sample standard deviation of 501.9. The estimated state distributions are compared to the simulated distributions in Figs. 8a and 8b. From these figures, it is apparent that both estimates closely approximate the original simulated distributions, yet depart from the probability plots of the original simulations most distinctly in their tails. The estimated Wet state has an approximate lognormal distribution, however there is a slight tendency for the Dry distribution to be heavier in the upper tail than the simulated Gaussian distribution.

The posterior distributions for the estimates of P_{WD} and P_{DW} had 90% credibility intervals of [0.186, 0.294] and [0.104, 0.163] respectively, both of which include the values used in the simulation. The Gaussian HMM was also fitted to this series, and the corresponding credibility intervals for P_{WD} and P_{DW} are [0.219, 0.299] and [0.200, 0.282] respectively. The second interval is quite misleading due to the conflicting requirement to approximate the marginal distribution by a mixture of normal distributions.

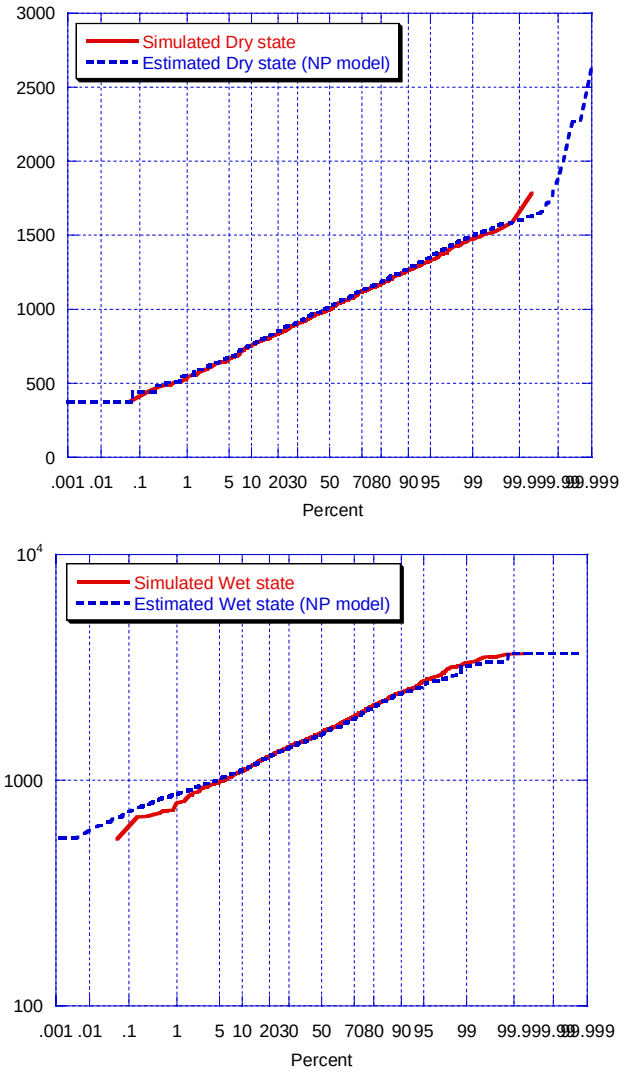

Fig. 8. *Simulated Dry state conditional Gaussian distribution (a), and simulated Wet state conditional lognormal distribution (b), of a variable (solid line) compared with the estimated distributions (dotted lines) from the NPHMM*

MARKOV CHAIN SIMULATED FROM TWO POISSON DISTRIBUTIONS

In this fourth example, a discrete-valued Markov chain of length 1000 was simulated with $P_{WD} = 0.25$ and $P_{DW} = 0.15$, with random samples for a Wet state being simulated from a Poisson distribution with a mean 15, and the Dry state simulated from a Poisson distribution with mean of 8. A histogram showing the marginal distribution of the simulated Markov chain is shown in Fig. 9. After fitting the two-point

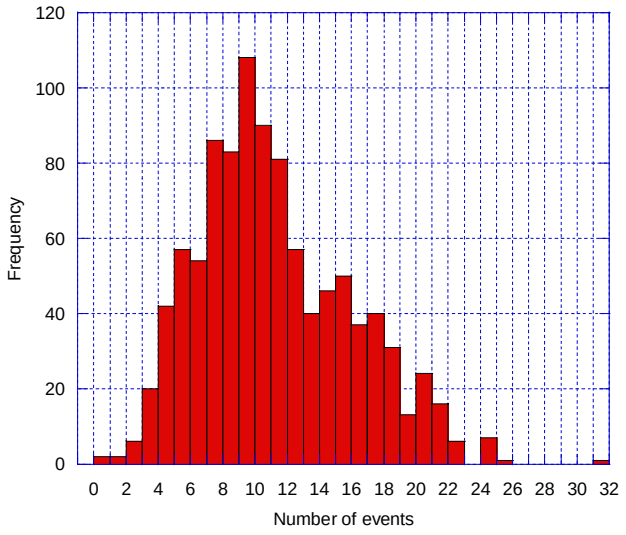

Fig. 9. *Marginal distribution of the Markov chain with two state conditional Poisson distributions*

NPHMM, Wet and Dry state distributions are estimated with respective means 15.324 and 8.219 and variances 13.525 and 9.033. The estimates of the variance are consistent with a Poisson distribution for which the mean equals the variance. The posterior distributions of the transition probabilities P_{WD} and P_{DW} have means 0.228 and 0.128 respectively. The 90% credibility intervals for these transition probabilities are [0.177, 0.281] and [0.096, 0.164], both of which include the values used in simulation. For comparison, a HMM with Poisson distributions describing each state was fitted, and the corresponding credibility intervals for P_{WD} and P_{DW} are [0.191, 0.289] and [0.111, 0.172]. These intervals also include the values used in simulation, and are somewhat narrower, again expected due to the assumptions of the parametric model being valid. In most instances however, realistic models for the underlying distributions will not be known.

Sensitivity of the NPHMM to length of input data

In the previous section, the NPHMM was fitted to Markov chains that were simulated over a length of 1000. However, if the NPHMM was employed to identify climate states within time series of annual rainfall, there will often be between 50 and 100 data points from which to observe hidden state transitions. By simulating two-state Markov chains over lengths of 50, 100, 500, 1000 and 10 000, the influence of input length on parameter estimates is investigated. Markov chains are simulated with transition probabilities $P_{WD} = 0.25$ and $P_{DW} = 0.15$, with random samples for the Wet state drawn from $N(2000,300^2)$ and the Dry state from $N(1500,300^2)$. These parameters were chosen to produce two distributions that overlap to a greater extent than the two Gaussian distributions used for earlier numerical simulation.

The distribution of the estimates of $P_{WD} + P_{DW}$ over the range of simulation lengths can be used as a measure of the efficiency of the NPHMM. The NPHMM is fitted to a single Markov chain of each length, and Fig. 10 shows the mean and 90% credibility interval for each posterior distribution, which can be compared with the true value of 0.4 for $P_{WD} + P_{DW}$. With the time series of length 50, the 90% credibility interval for $P_{WD} + P_{DW}$ is [0.276, 1.412]. This was followed up by 100 simulations of samples of length 50 and construction of the 90% credibility intervals. Ninety percent of these intervals included the true value of 0.4 but the remaining 10% all had a lower limit above 0.4, suggesting that the intervals are biased towards 1. This is reinforced by the result that 80% of these intervals included 1. It follows that, with small samples, the NP method is unlikely to distinguish a mixture from transitions between states with probabilities used in this example. This result has major implications for the use of this model to identify correct state distributions within time series of similar length to annual rainfall data. However, Fig. 10 suggests that the model can dramatically improve its estimates of transition probabilities when the length of input data is increased to 100.

To investigate further the parameter estimates made by the NPHMM, 100 Markov chains were simulated for each length analysed in Fig. 10, together with lengths of 40, 30, 20 and 10, with the two-point NPHMM fitted to each. Figures 11a and 11b show the variation in the estimates for the mean and standard deviation of the Wet state respectively after fitting the two-point NPHMM to Markov chains of

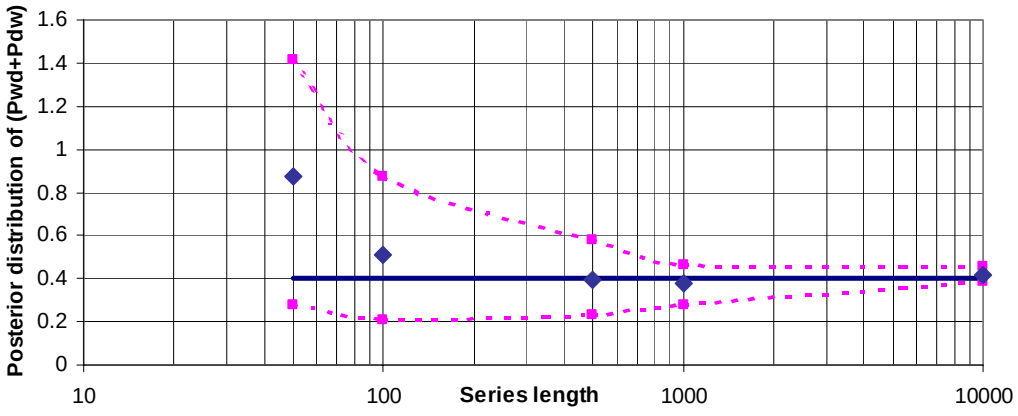

Fig. 10. *Mean and 90% credibility intervals for the posterior distribution for (P_{WD}+P_{DW}) from fitting the NPHMM to Markov chains of various lengths*

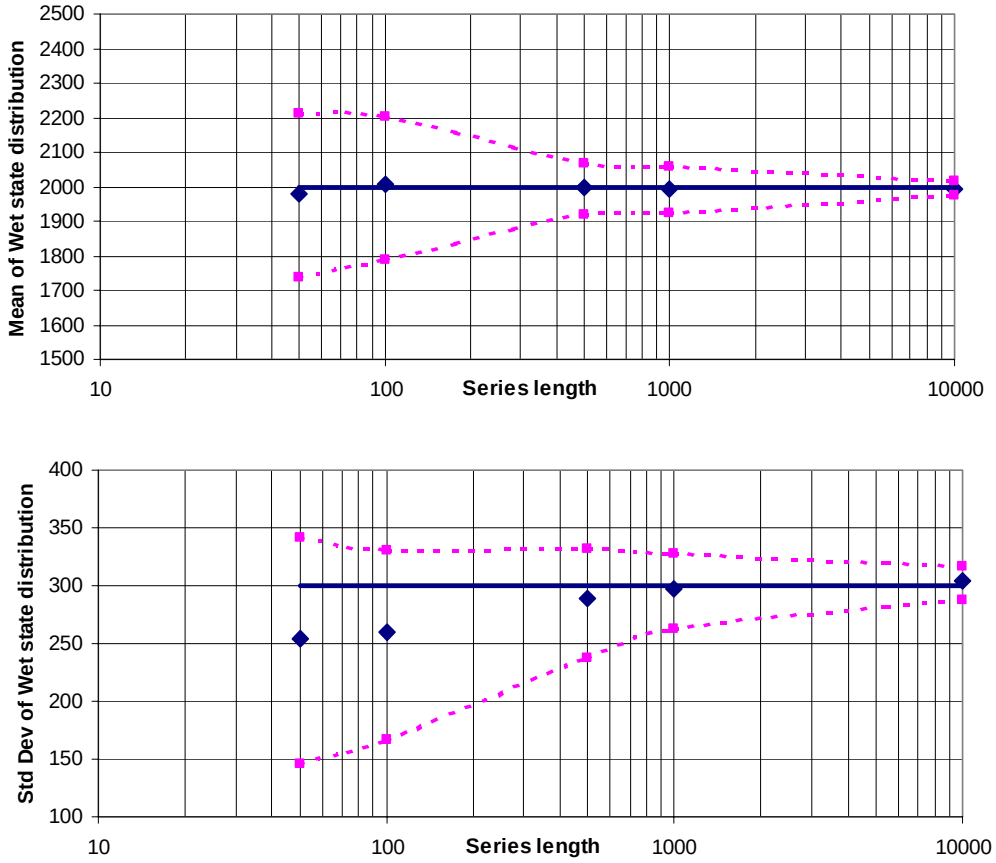

Fig. 11. *Mean (a), and standard deviation (b), of the Wet state distribution from fitting the NPHMM to Markov chains of various lengths. The means of 100 simulations are shown as diamonds, and limits within which 90% of estimates lie are shown by squares. The true value is 2000 for (a) and 300 for (b), both shown as solid lines*

various lengths. The solid lines indicate the average values for the Wet state mean and Wet state standard deviation respectively, from multiple simulations, with 90% of simulations lying within the dotted lines. Figure 11a illustrates that on average, estimates of the Wet state mean are close to the true value over all lengths; the uncertainty around these estimates increases dramatically for shorter input sequence lengths. From Figure 11b, it appears that the Wet state standard deviation tends to be underestimated for shorter sample lengths. This bias decreases as the sample size increases as does the variability of the estimator. Similar results were obtained for the Dry state.

Figure 12 shows the variation in the mean of the posterior distributions for estimates of $P_{WD} + P_{DW}$ from these 100 simulations at each model length. The solid line again indicates the average values, with 90% of the posterior means lying within the dotted lines. Although the means of the estimates of $P_{WD} + P_{DW}$ are close to the true value (0.40) at lengths of 500 and above, the estimates show considerable variability and a bias towards 1 for shorter sequence lengths. An explanation for this bias is that the uniform $U[0,1]$ priors for P_{WD} and P_{DW} imply a triangular distribution for their sum with a maximum value at unity.

The implication of this result is that, when applied to short annual rainfall series, a two-state NPHMM may struggle to provide accurate estimates of transition probabilities, thus impeding its use for simulation. At a length of 1000, approximately the length of monthly rainfall time series, the credibility intervals around transition probability estimates decrease dramatically, and at lengths of 10 000, the two-point NPHMM can identify both probabilities precisely.

Application of the NPHMM to observed data

The NPHMM is now applied to observed time series. In such series, the NPHMM is able to provide estimates for the unknown conditional state distributions and related transition probabilities. The distribution of the sum of

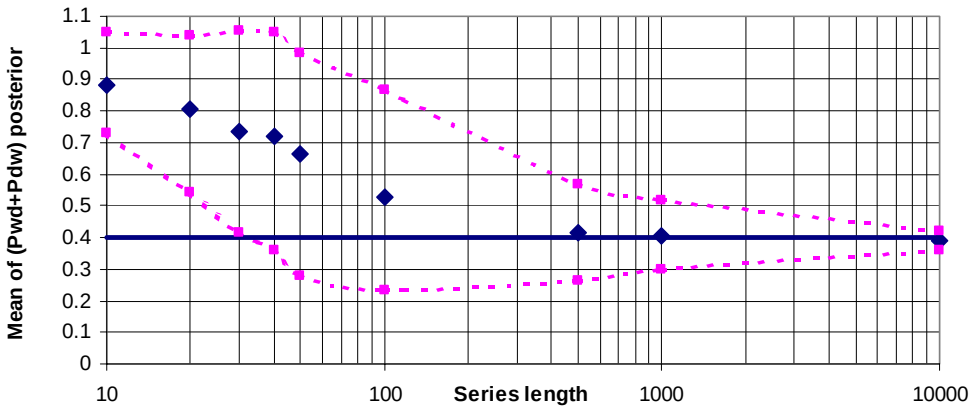

Fig. 12. *Mean of posterior distributions for estimates of* $(P_{WD}+P_{DW})$ *from fitting the NPHMM to 100 simulations of Markov chains at various lengths. The means of 100 simulations are shown as diamonds, and limits within which 90% of estimates lie are shown by squares. The true value is 0.4, shown as solid line*

transition probabilities is used as a test of the hypothesis that the fit of the HMM to rainfall and streamflow data is indicative of the influence of distinct climate states.

RAINFALL TIME SERIES

The NPHMM is fitted to the deseasonalised monthly rainfall records for ten cities in New South Wales, Australia, in order to estimate transitions between hidden states. The rainfall recorded at five cities (Port Macquarie, Newcastle, Sydney, Wollongong, and Batemans Bay) located on the coast of New South Wales, and at five sites located inland from each of these cities (Gunnedah, Dubbo, Orange, West Wyalong and Wagga Wagga) are analysed. The locations of these cities are shown in Fig. 13 with some statistics of the observed monthly rainfall series for each site presented in Table 1.

The sporadic tablelands and mountains that line the east coast of Australia known as the Great Dividing Range are shaded in Fig. 13. As the country's sole highland region, the Range has a strong influence upon local weather patterns, such that the regions to the west and in Australia's interior have a considerably lower rainfall regime than coastal sites. The statistics of rainfall shown in Table 1 support this point, with the five coastal sites showing higher mean monthly totals than the corresponding inland sites. This analysis can therefore assist in establishing whether coastal meteorological conditions influence estimates for HMM transition probabilities.

The Baum-Welch forward and backward recursion (as discussed by Bengio, 1999) can be included in the likelihood function of the NPHMM to obtain the hidden state probability time series. This time series is the posterior

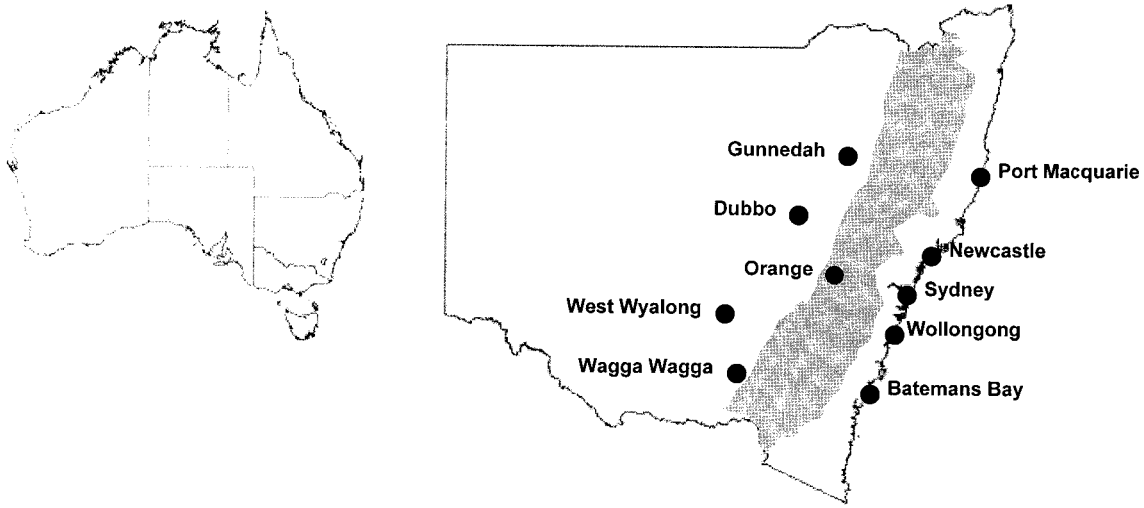

Fig. 13. *Locations of the ten rainfall sites in New South Wales, Australia that are used in analysis*

Table 1. Statistics of the observed monthly rainfall series

	Bureau of Meteorology rain gauge number	*Length (months)*	*Mean (mm)*	*Standard Deviation (mm)*	*Skew (mm)*	*1st Quartile (mm)*	*3rd Quartile*
Port Macquarie	60026	1572	126.92	107.99	1.92	49.23	177.35
Newcastle	61055	1608	94.70	78.22	1.88	40.80	125.18
Sydney	66062	1692	102.12	93.39	1.92	37.03	137.17
Wollongong	68069	888	94.87	99.17	2.46	30.60	126.35
Batemans Bay	69001	1200	84.29	90.31	2.74	25.13	110.80
Gunnedah	55023	1308	50.93	43.85	1.70	18.30	71.33
Dubbo	65012	1499	49.16	42.38	1.68	17.80	68.60
Orange	63066	1020	73.43	53.02	1.14	34.00	103.35
West Wyalong	50044	1104	39.26	33.91	1.72	15.00	55.05
Wagga Wagga	72151	1260	46.32	35.22	1.25	19.60	64.30

probability that each observation is derived from a particular state, which is defined for a Wet state as $p(s_t = W|Y_n)$ where n is the total number of observations. Values of $p(s_t = W|Y_n)$ closer to 0.5 indicate that the observation at time t is equally likely to be in either state, thus realising a weaker state signal. Thyer and Kuczera (2000) developed a state signal index (SSI) to measure how well the Wet and Dry state probability series are identified after fitting a two-state HMM. The SSI averages the $p(s_t = W|Y_n)$ values over the entire series, being calculated as follows:

$$SSI = \frac{\sum_{t=1}^{N} |p(s_t = W|Y_n) - 0.5|}{n} \quad (18)$$

If a higher number of data has values of $p(s_t = W|Y_n)$ closer to 0 or 1 then $SSI \to 0.5$. Therefore, higher SSI values indicate a better-defined state series. As a result, the SSI can be used as a tool to compare the state identification resulting from fitting the NPHMM to various time series.

The time series of observed monthly rainfalls are deseasonalised prior to analysis, such that each month has a mean of 0 and a standard deviation of 1. The two-point NPHMM was then fitted to each deseasonalised monthly rainfall series. Table 2 shows the mean and standard deviation of the posterior distributions for the two transition probabilities and for the sum of the two probabilities gained from fitting the NPHMM. From Table 2, it is apparent that the identification of hidden states is consistent in each of these monthly rainfall series, as 90% credibility intervals around the posterior means of the sum of the transition probabilities do not include 1 for any of the sites. In addition, the mean of the posterior distribution for the estimate of

Table 2. Mean and standard deviation of posterior distributions for transition probability estimates and SSI values from fitting NPHMM to deseasonalised (by monthly standardisation) rainfall series

	P_{WD}	P_{DW}	$(P_{WD} + P_{DW})$	*SSI*
Port Macquarie	0.256 (0.056)	0.315 (0.076)	0.571 (0.081)	0.3377
Newcastle	0.262 (0.066)	0.371 (0.142)	0.633 (0.149)	0.3686
Sydney	0.302 (0.060)	0.421 (0.071)	0.723 (0.075)	0.3543
Wollongong	0.260 (0.062)	0.376 (0.080)	0.637 (0.088)	0.3581
Batemans Bay	0.294 (0.079)	0.286 (0.099)	0.580 (0.131)	0.2604
Gunnedah	0.301 (0.094)	0.238 (0.081)	0.539 (0.128)	0.3197
Dubbo	0.221 (0.066)	0.178 (0.083)	0.399 (0.119)	0.2994
Orange	0.251 (0.075)	0.229 (0.058)	0.480 (0.103)	0.3574
West Wyalong	0.204 (0.093)	0.243 (0.083)	0.447 (0.123)	0.2674
Wagga Wagga	0.175 (0.063)	0.184 (0.069)	0.359 (0.102)	0.2852

$P_{WD} + P_{DW}$ at each of the inland sites is lower than the mean of the corresponding coastal site. This suggests that a two-state persistence effect may be more prominent in the rainfall observations of inland regions.

Table 2 also includes values of the SSI obtained from fitting the NPHMM to each of the deseasonalised monthly rainfall time series. The site producing the highest SSI value (0.3686) from the ten series analysed was Newcastle, with values ranging down to 0.2604 at Batemans Bay. The SSI value at three of the coastal sites is greater than the value of each corresponding inland site. Indeed, for the five pairs of sites analysed, the mean SSI values were higher for the coastal sites (0.336 to 0.305). However, there is no substantial evidence for any difference in the SSI between coastal and inland sites in the underlying population, the apparent difference being easily attributed to chance.

When interpreting the results from fitting the NPHMM to observed time series, it is important to also analyse the separation of the state distributions. If the NPHMM identifies precise values for transition probabilities, yet estimates the Wet and Dry states to have similar distributions, less physical significance can be gained from these results than if these distributions are more distinct. Table 3 shows some statistics of the estimated state distributions after fitting the NPHMM to the deseasonalised monthly rainfall series of four selected sites. It is apparent that for three of these sites, the first quartile of the Wet state lies above the third quartile of the Dry state, indicating that the two states are well separated. Although this separation is not achieved in the Gunnedah rainfall series, the distributions are still far enough apart to justify the choice of a two-state model.

The suitability of the NPHMM to simulate observed rainfall time series is shown in results from fitting the model to the deseasonalised Sydney monthly rainfall. For this example, the deseasonalised rainfall series is scaled such that each calendar month has a mean and standard deviation equal to that of the observed January record, a procedure that generates a time series of positive values. Figure 14

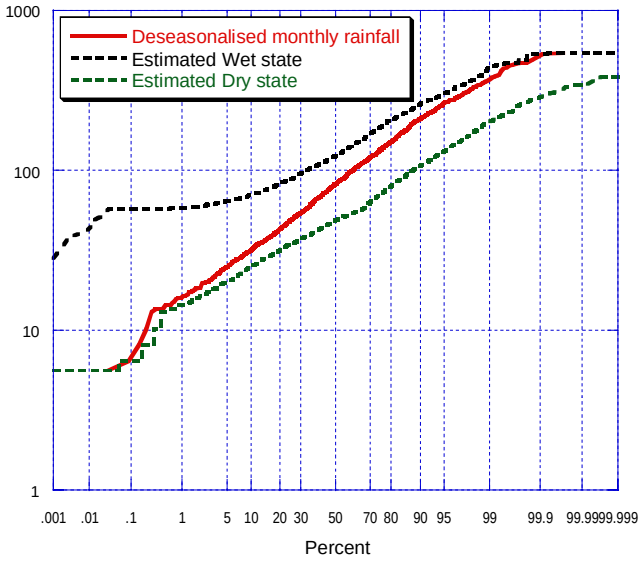

Fig. 14. *Lognormal probability plot showing the marginal distribution of the scaled deseasonalised (by monthly standardisation) rainfall series from Sydney (solid line) together with the estimated state conditional distributions in the Wet and Dry states (dotted lines)*

Table 3. Statistics of the estimated state distributions from fitting NPHMM to deseasonalised (by monthly standardisation) monthly rainfall data at selected sites

	Mean of estimated distributions	*Standard Deviation of estimated distributions*	*Skew of estimated distributions*	*1st Quartile*	*3rd Quartile*
SYDNEY					
Wet state	0.5634	1.0260	1.5059	-0.1989	1.0479
Dry state	-0.5863	0.4850	2.1078	-0.8923	-0.4281
ORANGE					
Wet state	0.6809	0.9839	0.7243	-0.0651	1.1760
Dry state	-0.5509	0.5492	0.7646	-0.9604	-0.1991
PORT MACQUARIE					
Wet state	0.3007	1.0129	1.8101	-0.3980	0.7467
Dry state	-0.6884	0.4252	1.6875	-0.9739	-0.5065
GUNNEDAH					
Wet state	0.6923	1.1012	0.9210	-0.1049	1.2645
Dry state	-0.2747	0.7858	1.1396	-0.8729	0.1619

shows this scaled deseasonalised record on a lognormal probability plot, together with the estimated Wet and Dry distributions from fitting the two-point NPHMM. This figure suggests that the approximate lognormal distribution of the monthly record is replicated in the two state distributions. Subsequently, a parametric two-state HMM was fitted to the data, with each state being sampled from lognormal distributions. This resulted in the posterior distributions for P_{WD} and P_{DW} having mean values of 0.333 and 0.387 with standard deviations 0.083 and 0.086 respectively. These estimates are in close agreement with those from the two-point NPHMM. This demonstrates that the NPHMM may be used to identify suitable parametric models for the simulation of conditional state distributions if that is desired.

STREAMFLOW TIME SERIES

The two-point NPHMM was also applied to streamflow time series, from the River Nile and three Australian rivers: River Murray, River Darling and Cooper Creek (see Table 4). The former series comprises the yearly minimal water levels in the Nile for the years 622–1281, measured at Roda Gauge near Cairo, as provided by Tousson (1925). Historically, this time series has been of specific interest, as it was amongst various geophysical time series whose analysis led to the discovery of the Hurst effect of long memory (Hurst, 1951).

Cooper Creek (length 1523 km) has a high hydrological variability, as shown by the high positive skewness in Table 4, and is part of the Lake Eyre Basin which covers 1.14 million km^2 of eastern inland Australia (Puckridge *et al.*, 2000). Mean daily flows are recorded at Cullyamurra gauging station, South Australia, over the period October 1973 to September 2002. Monthly totals used in this analysis are first standardised to make flows in each calendar month have a zero mean and standard deviation of unity.

The Murray-Darling Basin (MDB) drains 1.073 million km^2 (14%) of Australia, and includes the longest rivers, namely the Darling, rising in sub-tropical Queensland, and the Murray, rising in south-eastern New South Wales. Some 80-90% of the basin is arid or semi-arid, and most of the runoff is generated from rainfall and snowmelt in eastern catchments. Total runoff is merely 4% of annual rainfall. Estimates of natural flows in the Murray and the Darling were obtained from the Murray-Darling Basin Commission's (MDBC) Monthly Simulation Model (MSM). Flows in the MDB are calculated by the addition of observable flow and diversions and losses associated with upstream storages. Estimates of natural flows are then obtained by setting diversions from the system and the storages to zero. Total flows are estimated at Lock 10, immediately downstream of the junction between the Darling and Murray at Wentworth, New South Wales, and Darling flows are estimated at Burtundy, on the lower Darling. Murray flows above Lock 10 were estimated by subtracting Darling flow from total Murray-Darling flow at Wentworth. Time series of such reconstituted natural monthly flows in both the Murray and the Darling for January 1892 to December 1999 were analysed.

The seasonal components in the time series of monthly flows for the Murray, Darling and Cooper were removed in a similar process as used in the analysis of monthly rainfall series. However, these deseasonalised series were then scaled so that each calendar month had a mean and standard deviation equal to the observed data in a nominated month that produced a time series of positive values.

The two-point NPHMM was fitted to the four time series and the mean and standard deviation of the posterior estimates for the two transition probabilities are displayed in Table 5. This table indicates that the NPHMM can identify two distinct transition probabilities with reasonable precision in each series. Furthermore, the values of the transition probabilities are low, suggesting that the two-state HMM is able to provide a good representation of the variability in each of these time series by estimating long sojourn times in each climate state. Table 5 also includes values for the SSI obtained from fitting the NPHMM to each of the streamflow time series. The SSI values for the streamflow records are all close to 0.5, which suggests that the HMM is able to identify strong Wet and Dry distributions in each time series. These SSI values are significantly higher than values from the deseasonalised rainfall records analysed. If

Table 4. Statistics of the selected streamflow time series

	Length	*Mean*	*Standard Deviation*	*Skew*	*1st Quartile*	*3rd Quartile*
Nile minimum annual level	663 years	1148.1	88.7	0.244	1094.0	1205.0
Murray natural monthly flows	1296 months	1082.6 GL	1105.5 GL	2.245	332.9 GL	1384.4 GL
Darling natural monthly flows	1296 months	166.8 GL	208.6 GL	2.273	20.4 GL	243.8 GL
Cooper observed monthly flows	348 months	136.4 GL	669.3 GL	9.589	0.006 GL	36.172 GL

Table 5. Mean and standard deviation of posterior distributions for estimates of transition probabilities, and SSI values, from fitting NPHMM to streamflow data

	P_{WD}	P_{DW}	$(P_{WD} + P_{DW})$	SSI
Nile	0.121 (0.028)	0.067 (0.018)	0.188 (0.042)	0.4431
Murray	0.085 (0.010)	0.092 (0.010)	0.177 (0.018)	0.4688
Darling	0.117 (0.012)	0.109 (0.010)	0.226 (0.021)	0.4620
Cooper	0.241 (0.039)	0.114 (0.020)	0.355 (0.054)	0.4876

the concept of climate states is accepted, their influence can be identified better in streamflow data than in rainfall.

Table 6 shows some statistics of the estimated Wet and Dry state distributions gained from fitting the NPHMM to each of the streamflow time series. For each series, the means of the two state distributions are well separated, and with the first quartile of each Wet state being greater than the third quartile of the corresponding Dry state, the NPHMM can identify distinct state distributions in each time series. With low transition probabilities, high SSI values and state distributions that are well separated, the two-state structure of the HMM appears to be justified.

Discussion

The method introduced to fit a non-parametric (NP) HMM to hydrological data is an approach to directly identify modes of variation consistent with climatic influence. The assumption of a 'hidden' climatic indicator oscillating between only two states is a simplification of large-scale atmospheric circulation, yet is consistent with previous research (e.g .Thyer and Kuczera, 2000; Katz and Zheng, 1999). Thyer and Kuczera (2000) argued that realisations of Gaussian HMMs can be distinguished from AR(1) models, with the former being more realistic for simulating Australian hydrological data. However, HMMs with unrealistic conditional distributions will produce biased estimates for transition probabilities and therefore will not provide realistic simulations.

Hydrological data such as monthly rainfall time series are often skewed, and therefore any methods employed to simulate such series must have the ability to replicate this feature. Unrealistic assumptions about the form of state conditional distributions, such as fitting Gaussian HMMs when the distributions are non-Gaussian, will force biased transition probabilities due to the conflicting requirement to model the skewed marginal distribution. The use of the NPHMM, which makes no assumptions about climatic state

Table 6. Statistics of the estimated state distributions from fitting NPHMM to the observed Nile record and to the scaled deseasonalised (by monthly standardisation) streamflow records from the Murray, Darling and Cooper

	Mean	*Standard Deviation*	*Skew*	*1st Quartile*	*3rd Quartile*
Nile					
Estimated Wet state	1190.2	71.1	0.529	1142	1241
Estimated Dry state	1070.7	58.8	0.253	1034	1103
Murray					
Standardised to February record	384.89 GL	234.66 GL	2.048	230.27 GL	470.65 GL
Estimated Wet state	535.37 GL	233.68 GL	2.042	377.60 GL	622.90 GL
Estimated Dry state	224.05 GL	58.81 GL	-0.334	187.20 GL	262.21 GL
Darling					
Standardised to March record	15.714 GL	15.951 GL	2.099	4.911 GL	21.982 GL
Estimated Wet state	26.465 GL	16.383 GL	1.821	14.912 GL	34.986 GL
Estimated Dry state	5.147 GL	2.775 GL	0.676	3.236 GL	6.563 GL
Cooper					
Standardised to April record	197.8 GL	397.5 GL	3.774	45.2 GL	122.4 GL
Estimated Wet state	449.08 GL	579.59 GL	1.926	117.24 GL	420.24 GL
Estimated Dry state	60.60 GL	31.00 GL	0.326	29.51 GL	89.04 GL

distributions, will have a clear advantage in the identification of these distributions and also for the simulation of hydrological data. The possible loss of efficiency in using a non-parametric framework can be compensated for if the model is first used to identify suitable parametric families from which conditional state distributions can be simulated. The NPHMM was shown to recover model parameters successfully from time series generated from both continuous and discrete parametric families.

In the hydrological literature, persistence is generally defined as run times on either side of the long-term mean longer than would be expected for an independent process. This does not correspond to the asymptotic definition of a persistent stochastic process given by Beran (1994), for example, that the sum of autocovariances over all lags tends to infinity. The HMM is not persistent in this latter sense (Whiting *et al.*, 2003a). When the NPHMM was fitted to various rainfall and streamflow time series, the latter displayed longer sojourn times in the two states.

A limitation of the NPHMM is that the inverse of the mapping from the data to plotting positions, which is used in simulation, will be constrained to be within the range of the original data. This is unsatisfactory if the simulation is required to provide realistic extreme values. In these cases, the NPHMM can be used to identify realistic parametric forms for the state conditional distributions and a suitable parametric HMM can then be fitted.

Conclusions

A non-parametric hidden Markov model (NPHMM) that does not make assumptions about the form of conditional state distributions has been shown to be competitive with parametric HMMs for identifying model parameters in several numerical simulations. Parametric HMMs that make inappropriate assumptions about the form of underlying state distributions have been shown to suffer from their requirement to produce a mixture of distributions to match the marginal distribution. This can bias the estimation of transition probabilities, whereas the NPHMM avoids this potential source of estimation error. The non-parametric approach identifies statistically significant evidence for hydrological persistence in both monthly rainfall and streamflow records, providing a sufficient length of data is available for analysis.

Acknowledgements

The authors wish to thank Associate Professor George Kuczera for several in-depth discussions on this work. This project was supported by an Australian Research Council Discovery grant.

References

Bengio, Y., 1999. Markovian Models for Sequential Data. *Neural Computing Surveys* **2**, 129–162.

Beran, J., 1994. *Statistics for Long-Memory Processes*. Chapman & Hall, New York, USA.

Brooks, S., 1998. Markov Chain Monte Carlo method and its application. *The Statistician,* **47**, 69–100.

Chib, S., 1995. Marginal Likelihood From the Gibbs Output. *J. Amer. Stat.Assn.,* **90**, 1313–1321.

Churchill, G.A., 1989. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.,* **51**, 79–94.

Ephraim, Y. and Merhav, N., 2002. Hidden Markov Processes. *IEEE Trans. Information Theory,* **48**, 1518–1569.

Hughes, J.P. and Guttorp, P., 1994. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.* **30**, 1535–1546.

Hurst, H.E., 1951. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Engrs*, **116**, 770–799.

Juang, B.H. and Rabiner, L.R., 1991. Hidden Markov Models for Speech Recognition. *Technometrics,* **33**, 251–272.

Katz, R.W. and Zheng, X., 1999. Mixture model for overdispersion of precipitation. *J. Climate,* **12**, 2528–2537.

Kuczera, G. and Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *J. Hydrol.,* **211**, 69–85.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N.,Teller, A.H. and Teller, E., 1953. Equations of state calculations by fast computing machines. *J. Chem. Physics,* **21**, 1087–1091.

Puckridge, J.T., Walker, K.F. and Costelloe, J.F., 2000. Hydrological persistence and the ecology of dryland rivers. *Regulated Rivers: Res. Manage.*, **16**, 385–402.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Robert, C. P., Ryden, T. and Titterington, D.M., 2000. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. Roy. Stat. Soc., Series B,* **62**, 57–75.

Thyer, M.A., 2000. *Modelling long-term persistence in hydrological time series*. Ph.D Thesis. Department of Civil, Surveying and Environmental Engineering, University of Newcastle, Australia.

Thyer, M.A. and Kuczera, G.A., 2000. Modelling long-term persistence in hydroclimatic time series using a hidden state Markov model. *Water Resour. Res.,* **36**, 3301–3310.

Tousson, O., 1925. *Mémoire sur l'Histoire du Nil,* Mémoires de l'Institut d'Egypte.

Whiting, J.P., Lambert.M.F. and Metcalfe, A.V., 2003a. Modelling persistence in annual Australian point rainfall. *Hydrol. Earth Syst. Sci.,* **7**, 197–211.

Whiting, J., Lambert, M., Metcalfe. A. and Walker, K.F., 2003b. Searching for persistence in unregulated River Murray streamflows. *28th Int. Hydrology and Water Resources Symposium*, 10-14 November, Wollongong, Australia.

Zucchini, W. and Guttorp, P., 1991. A Hidden Markov Model for Space-Time Precipitation. *Water Resour. Res.,* **27**, 1917–1923.